DATA MANAGEMENT IN THE LOANSTAR PROGRAM

Robert E. López and Jeff S. Haberl
Energy Systems Laboratory
Department of Mechanical Engineering
Texas A&M University
College Station, Texas

ABSTRACT

This paper discusses the complexity of managing building energy usage data for many buildings. The history and methodology of data collection at the Texas LoanSTAR Monitoring and Analysis Program, a large multimillion dollar project, is given as an example. The differences in methodology of managing data for one/several buildings versus many buildings are given and discussed. Primary database design and quality assurance issues that should be thought out at the beginning of any large project of this type are given. The importance of intergroup communication throughout the project is stressed. Current areas of software development are discussed in detail, followed by future directions for the project.

INTRODUCTION

The Texas LoanSTAR (Loans to Save Taxes And Resources) Program was established in 1988 by the State of Texas Governor's Energy Office. This eight year, $98.6 million program uses revolving loans to fund energy conserving retrofits in state and local government buildings. Established the following year, the Monitoring and Analysis Program (MAP), attempts to measure and report energy savings from the various retrofits. Several months prior to installing retrofit equipment in the larger state agencies, data acquisition systems are installed to monitor energy consumption. By analyzing data collected both before and after the retrofit, the overall effectiveness of the energy conservation measures can be determined.

Since 1989, the number of buildings which are being monitored, and consequently the amount of data which is collected, has increased in a dramatic fashion. The first year of the project saw hourly data collection from one local building in College Station. Since October of 1990, over 50 buildings from around the state have been added at irregular intervals. The first set of buildings to be added were buildings at the University of Texas at Austin (12 buildings), followed immediately by the State Capitol Complex (9 buildings) in Austin. After this initial rush of LoanSTAR Program buildings, collection of hourly National Weather Service data from locations throughout Texas began in November 1990. Currently, the LoanSTAR MAP is collecting data from 54 buildings located in various cities throughout Texas as well as weather data from 50 National Weather Service stations. Figure 1 gives an historical perspective of the data collected for the MAP, increasing from the single prototype site at the main Texas A&M campus to the 54 sites across the state. From these

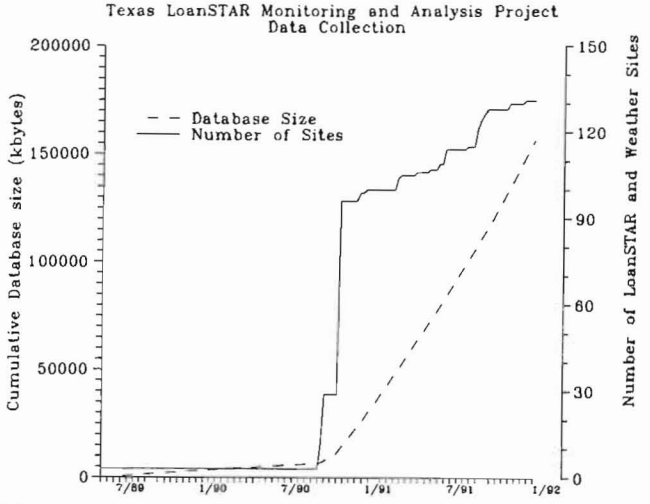

***Figure 1**: Data Collection at the LoanSTAR MAP.* This graph shows the increase in the number of sites and the amount of data in the database from the beginning of the project in May 1989 to the time this paper was written, January 1992.

sites and the National Weather Service stations over 1700 different data channels are extracted; every week the size of the MAP data archive increases by nearly three megabytes. As new sites are added almost monthly, the management of the data is of prime importance.

LoanSTAR MAP data management includes five major functions: polling the data from the data acquisition systems (DAS), processing the data from all the various sites into a reasonably generic format, controlling data quality, generating reports, and retrieving data for analysis. Polling, processing, and report generation have been discussed previously (López and Haberl, 1992; Claridge et al., 1991; Haberl et al. 1990a,b). Following a brief overview of these areas, this paper describes the methods employed at LoanSTAR for managing the data archive and briefly explains some of the issues that have been faced in the management of building energy data for multiple buildings.

OVERVIEW OF DATA MANAGEMENT IN THE MAP

Polling (Data Collection)

The first step in this data management scheme is the entry of data into the system. This includes both the retrieval of building data from the remote data loggers and the acquisition of National Weather Service data. At LoanSTAR, each building is polled weekly. Using IBM PC-based communication software supplied by the DAS manufacturers,

Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates, Dallas, TX, May 13-14, 1992

the LoanSTAR MAP currently downloads data as basic ASCII columnar text. Each remote DAS collects hourly consumption information which is stored in onboard volatile memory. Because these systems have a finite amount of memory, the DAS is polled once per week to avoid older data being overwritten by newer data and lost. As each site is polled, the dataset itself is saved as one text file per site per week. Therefore, every week 54 new raw data files containing 168 hourly records each are saved. Prior to processing, these files are stored in a temporary directory on the polling PC. After the data have been processed, the raw files are archived to tape.

The National Weather Service data records are initially collected by the Texas A&M Meteorology Department using satellite technology. The MAP has been allowed to transfer weather data over the campus Ethernet for internal use in analysis at minimal cost. The Meteorology Department is actually collecting data from NWS locations nationwide; therefore, some initial filtering is done on the Meteorology computer to extract only Texas sites before the package is transferred. This relieves the strain on the campus network and obviously reduces the disk space required on the LoanSTAR computer system. The NWS dataset requires substantial processing to be usable by the LoanSTAR analysis teams. As with the building data, the raw files are archived onto tape. In both cases, if problems are identified somewhere in the processing stream at a later date, all raw data are still readily available for reprocessing.

Processing and Quality Control

Processing of the weekly dataset and essential quality control are performed through a combination of public domain utilities, commercial software, and routines written in-house to knit the data streams together. It should be noted that a goal of the LoanSTAR MAP has been to use inexpensive existing software wherever possible to decrease development time. This methodology has caused an interesting assortment of modules to evolve over the past two years. The routines written by LoanSTAR are generally programmed in either GAWK or C++. GAWK (FSF, 1989) is a public domain version of AWK, a powerful UNIX file processing language. GAWK is available in both DOS and UNIX versions.

The cornerstone of the processing and quality control areas is a public domain DOS program called ARCHIVE, which was developed at Princeton University (Feuermann and Kempton, 1987). ARCHIVE is a general purpose program for manipulating and checking columnar data. Unfortunately, the program can process numeric data only. The software used to poll the loggers in the MAP place certain nonnumeric status fields in the data for diagnostic purposes. A filter was written to check and remove these status fields as well as to check and remove any off-hour readings. The filter keeps a log file for reporting offending status fields and off-hour readings. This log file can be used to determine power failures at the site.

As input, ARCHIVE requires the prefiltered data and another file containing a channel table, which is a description of the input columns and the output format, including any translations which might need to take place. Table 1 gives a short example of an ARCHIVE channel table. ARCHIVE allows several date translation schemes. For example, given a MM DD YY date and a time, ARCHIVE can report the corresponding decimal date. This type of translation is essential for graphing time series data for periods of varying length. ARCHIVE output is the standard LoanSTAR format. For each hourly record (row) in a file, this format is as follows:

site#	MM	DD	YY	JulianDate	DecimalDate	Time	data1	...	datan
100	01	01	92	92001	4383.0000	000	21.5	...	46.2

The site number is a unique three digit LoanSTAR number that is associated with each site. These site numbers are

Table 1: *The ARCHIVE channel table for LoanSTAR site 100.* The channel descriptor table for a basic LoanSTAR site is given. ARCHIVE channel tables allow for a certain amount of quality control by associating static lower and upper bounds with each channel.

Date MM/DD/YY (YY DDD)	Time HH:mm	Raw-Data lin pos	coln pos	Arch coln pos	Name of Channel	Archive Units	Arch Format	Conv'n Code	Conv'n Constants	Error Code	Error Constants	Channel Description
#												
07/03/90	00:00	1	0	0	Begin	Education						Beginning date
07/03/90	00:00	1	1	1	Bldg. #	MM	I3	2	0 100	0		Building Number
07/03/90	00:00	1	1	2	Mon-Raw	MM	I3	1		0		Month
07/03/90	00:00	1	2	3	Mon-Raw	DD	I3	1		0		Day
07/03/90	00:00	1	3	4	Mon-Raw	YY	I3	1		0		Year
07/03/90	00:00	1	3	5	Greg-Jul	MMDDYY	I5	24	1 2	0		Gregorian Date to Julian
07/03/90	00:00	1	4	7	Time	HH mm	I5	16	5	0		Time
07/03/90	00:00	1	3	6	Greg-Dec	DDD.frac	F10.4	28		0		Gregorian Date to Jul.Decimal
07/03/90	00:00	1	6	8	RAFs I-1	F9.3	F9.3	1		1	-5 500	Return Air Fans I (kWh/h)
07/03/90	00:00	1	7	9	SFs I-1	F9.3	F9.3	1		1	-5 500	Supply Air Fans I (kWh/h)
07/03/90	00:00	1	8	10	RAFs II-1	F9.3	F9.3	1		1	-5 500	Return Air Fans II (kWh/h)
07/03/90	00:00	1	9	11	SFs II-1	F9.3	F9.3	1		1	-5 500	Supply Air Fans II (kWh/h)
07/03/90	00:00	1	10	12	S/R/5A-1	F9.3	F9.3	1		1	-5 500	Supply/Return-5A Individual (kWh/h)
07/03/90	00:00	1	11	13	S/Rs/MC-1	F9.3	F9.3	1		1	-5 500	Supply/Return-MC (kWh/h)
07/03/90	00:00	1	12	14	CHWP-1	F9.3	F9.3	1		1	-5 500	CHWP (kWh/h)
07/03/90	00:00	1	13	15	ChWP-1	F9.3	F9.3	1		1	-5 500	ChWP (kW)
07/03/90	00:00	1	14	16	kWh M1	F9.3	F9.3	1		1	0 99999	Bldg kWh Meter A (kWh/h)
07/03/90	00:00	1	15	17	kWh M2	F9.3	F9.3	1		1	0 99999	Bldg kWh Meter B (kWh/h)
07/03/90	00:00	1	16	18	CondensM	F9.3	F9.3	1		1	0 99999	Condensate Meter (gal)
07/03/90	00:00	1	17	19	ChWBtu	F9.3	F9.3	1		1	0 99999	ChW Btu (kBtu)
07/03/90	00:00	1	18	20	ChWFlow	F9.3	F9.3	1		1	0 99999	ChW Flow (gal)
03/11/99	23:00	1	0	0	End	Education						

extremely important in the quality control process. The first stage of any data audit involves checking the site numbers in the file against each other as well as the site number required by the analyst. A shortcoming of circulating ASCII data in a DOS environment is the ease with which data can become confused or overlaid among sites.

For additional quality control, static lower and upper bounds can be associated with any channel in the channel table. For readings outside the specified range, ARCHIVE will flag the value in a diagnostic log file as well as replace the suspect value with some predefined "bad data" marker in the output data. This allows an automated method for assuring that data are reasonable. For example, a dry bulb temperature channel for a site in most parts of Texas might have a lower bound of -10 and an upper bound of 120. Obviously, it is quite useful to have a program check the 168 hourly readings each week rather than doing this by hand. ARCHIVE can also perform simple data translations. For example, linear transformations can be used to convert between different units. In fact, a linear translation is used to attach the site number to every record in the output. ARCHIVE produces two files as output: a diagnostic log file and the actual output data. The log files are inspected every week to ensure data quality.

After passing through ARCHIVE, the data file is scanned for missing hours. It is not terribly uncommon for a data logger to lose power in the field. Usually these loggers have battery backups that perform a minimal amount of work: refreshing the internal memory. This allows the logger to keep any data it has collected up to this point, but the logger does not collect any new data until the power is restored. This in turn creates gaps in the dataset. For purposes of merging weather data and certain types of analysis, it was determined early in the project that missing hours should be added back into the data with a "bad data" marker inserted for all data values. Therefore, the data file is filtered through a generalized AWK script to scan for missing hours and put them back in as necessary. While the concept of scanning and replacing missing hours is easy to understand, a generalized program must take into account day boundaries, month boundaries, year boundaries, leap year boundaries, and several other special conditions. As with ARCHIVE, the output of this script consists of two files: a diagnostic log file which reports the number of hours added back in, and the actual output. This file is the final version in which all LoanSTAR hourly data is kept. These processed files are archived to two tapes and also transferred to the MAP's UNIX file server over the campus Ethernet. Storing the data on a large file server allows immediate access to all the data across all sites as well as providing access to powerful tools such as the UNIX version of AWK, C, C++, and commercial statistics packages such as SAS (SAS, 1990).

National Weather Service data are translated into the LoanSTAR format through the use of several UNIX shell scripts and supporting AWK scripts that convert the incoming data into a format readable by ARCHIVE. The weekly weather dataset is then processed in a fashion similar to LoanSTAR building data with one important difference: the weekly file contains the hourly records from all Texas weather sites. This was done to keep the overall processing scheme as efficient as possible. C++ routines are used to extract particular locations from the weekly dataset. As shown in Figure 2, this scheme has proven particularly useful for creating verification crossplots between LoanSTAR weather data and the National Weather Service data.

Weekly Report Generation

While simple automated quality control checks have been implemented, a key function of the whole process is the production of weekly verification plots. These plots are circulated between the project's Principal Investigators and research staff in a bundle referred to as the Inspection Plot Notebook (IPN). An example page of verification plots is given as Figure 3. The plots allow for possible problems with the data to be identified by visual checks. Graphical presentation of the data on a weekly basis adds tremendously to the quality control done by ARCHIVE and is much less time consuming than scanning the actual ASCII data columns. Because of the long stream of software filters that the data are subjected to prior to the production of the plots, any potential problems are usually brought to the attention of the Database Manager, who determines if a processing problem could have corrupted the data. If this is not the case, then a genuine data problem may have occurred (for example, a metering problem), and an appropriate message is forwarded to the field engineers.

These weekly plots are currently produced with a commercial graphics package, along with supporting AWK scripts and a controlling DOS batch file. Three different kinds

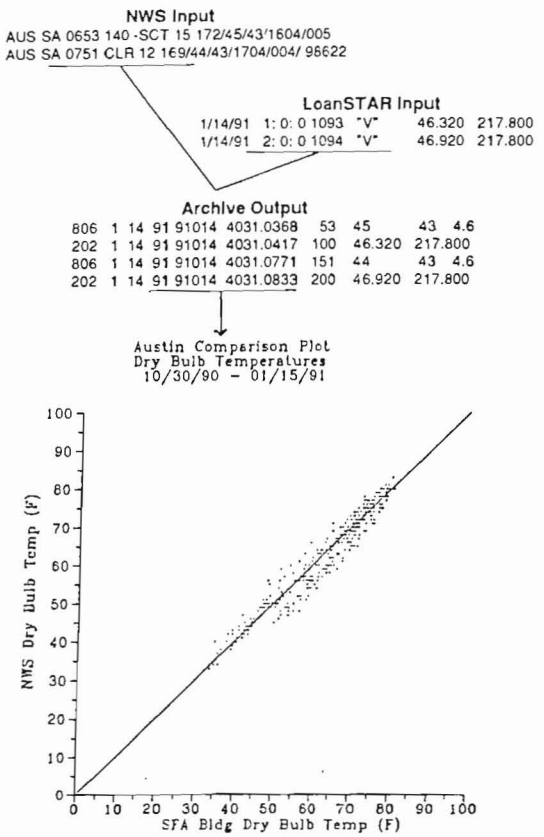

Figure 2: Combination of two data streams for verification. Data records from the National Weather Service and LoanSTAR weather sites can be combined into a standard format to create crossplots for data quality control.

of pages are created for the notebook: time series readings of all channels; summary pages that include scatterplots of some channels (thermal channels and motor control center electricity consumption) versus temperature and derived time series readings of the primary data types (whole building electricity consumption and thermal channels); and scatterplots of LoanSTAR weather data versus National Weather Service data (as in Figure 2). The batch files and supporting scripts allow these pages to be produced from the processed data on a weekly basis with a simple command. However, it should be noted here that since every site is different, the initial set up time is significant. The LoanSTAR MAP currently produces roughly 950 of these small graphs, 80 pages in all, each week. The start-up time was substantial and the computing time required each week is nontrivial. Additionally, the logistics of actually printing, copying, filing and routing the IPN each week should be addressed. All told, the entire process from polling through the circulation of plots requires between 20 and 30 man-hours every week. This does not include the time spent by the Principal Investigators and project staff reviewing the plots themselves. It is estimated that the MAP spends an additional 50 man-hours each week reviewing the IPN, although this value is hard to quantify because the interested parties have different methodologies and goals.

Monthly Report Generation

The weekly generation of inspection plots allows for internal verification of the monitoring systems. However, a primary goal of the project as a whole is the creation of Monthly Energy Consumption Reports (MECR) which are sent back to the facilities. This is the major feedback route from the MAP back to the participating agencies. The content and value of the MECR is discussed elsewhere (Claridge et al., 1992). Briefly, each building receives a six page report that consists of:

1) A title page with usage totals and comments to the agency.

2) Scatterplots of daily hot water usage versus average daily outside dry bulb temperature and daily chilled water usage versus average daily outside dry bulb temperature.

3) Time series plots of hourly hot water usage and chilled water usage.

4) Time series plots of total building electricity consumption and outside ambient conditions (dry bulb temperature and relative humidity) for this area.

5) Two three dimensional plots that graph hourly total building electric in daily slices.

6) A summary page with information about the building (square footage, occupancy schedules, retrofit dates, et cetera).

The plots for pages two, three, and four are produced with the same commercial graphics package used to produce the IPN, while the three dimensional plots on page five are created with SAS on the UNIX file server. The Postscript versions of the graphs are knitted together with the TEX typesetting program

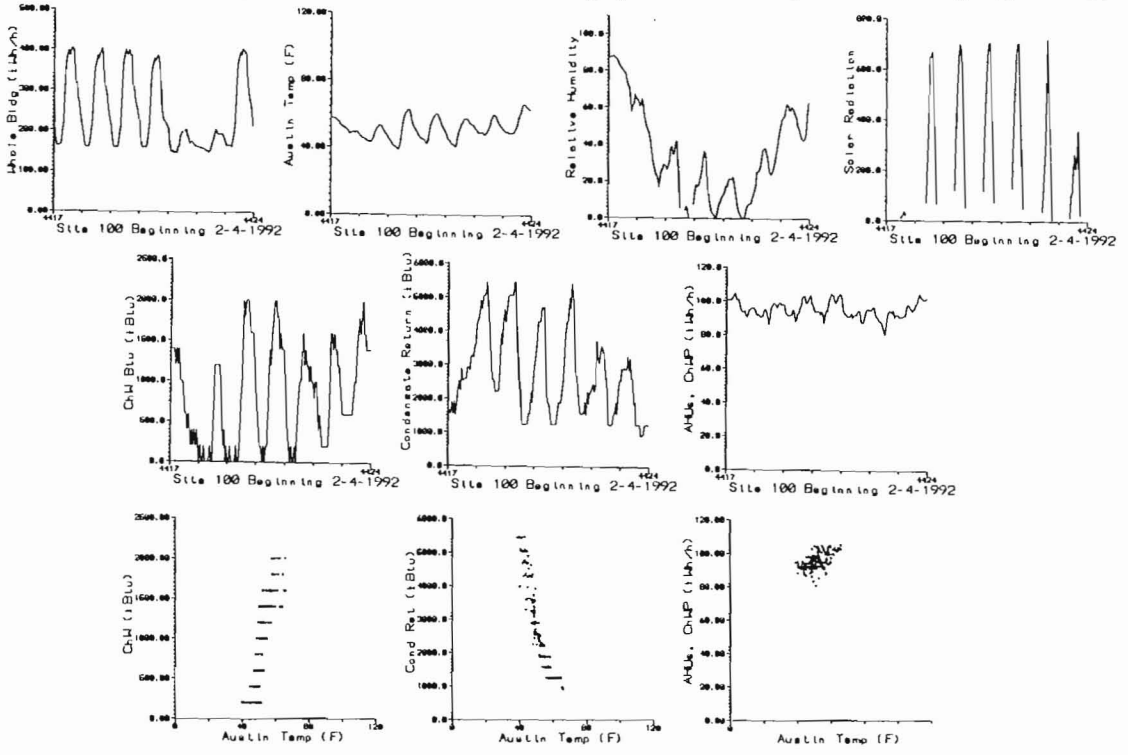

***Figure 3**: A typical page from the Inspection Plot Notebook (IPN).* The IPN consists of small graphs of each channel at all of the various LoanSTAR sites. These pages are circulated amongst the Principal Investigators and LoanSTAR staff for data quality control purposes.

(Knuth, 1986). Pages one and six are created entirely within TEX.

Although the data are kept on-line, a significant amount of work must be done each month to create quality reports for external distribution. Each month, the appropriate weekly files are concatenated into monthly files and a conversion script is used to convert the hourly data into daily total data for page two. Average daily dry bulb temperatures are calculated by the script using the available hourly weather data. After an initial set of plots is created, it is circulated to determine the comments for page one as well as to catch any processing errors.

After two rounds of circulation, the plots are sent to the corresponding agencies. Not including review time by the various Principal Investigators, the MECR requires roughly 120 man-hours to produce. At the current level, usually two people working full-time can produce the MECR in about a week and a half. Including another five days for comment circulation and printing, the target date for report delivery is somewhere around the 20th of the following month.

Data Retrieval and Delivery

Immediately prior to the production of the MECR, a monthly dataset containing the previous month's data for all sites, weather data for each region, and daily total data is produced for the analysis teams. Currently, a standard data update is provided on floppy disk, usually within two days after the final polling cycle of the previous month. The textual, columnar format of the data is generic enough to be used directly as input to most commercial statistics packages.

Additionally, a very impressive, commercially available data browsing package called VOYAGER (Lantern, 1990) is used by the analysis teams to view the data. This software was originally developed to analyze meteorological data, but has been adapted very well to the analysis of building energy data. As with the statistics packages, this program can read textual, columnar data readily. However, this particular package transforms the data into a compiled database for internal use. For agencies which have purchased this software, the LoanSTAR MAP provides monthly precompiled updates that allow the agencies to browse through their own consumption data.

DISCUSSION

Since the beginning of the MAP in early 1989, this project has been able to construct a valuable set of public domain tools and a process for collecting, verifying, and reporting monitored building energy data on a regular, timely basis. This process has been developed basically from the bottom up: abstract ideas were formalized and matched with available software components when possible. Tailor-made filters have been developed to knit the modules together into a continuous stream. After the first addition of new sites in October 1990, the project was able to start into a MECR production mode within six months, a very brief design phase given the concurrent influx of new sites (Figure 1).

Figure 4 summarizes the flow of information through the various routines and functions which make up data management at the LoanSTAR MAP. Although progress has been rapid, each phase of the data management scheme has contained its share of problems. At the beginning of the project, the problems were centered around polling and processing. Specifically, there were no hard and fast rules on how to poll or when to poll, much less any real direction on how to store the data. Although those questions were resolved early on, as the number of sites multiplied, new problems began to appear concerning quality control and report generation. Additionally, as more and more data are processed and stored, answers to old questions become obsolete and need to be rethought. Many problems seem to manifest themselves differently as a function of the number of sites being maintained. Empirical evidence from the project's experience has shown three distinct stages in the lifecycle of the project to date: maintenance of one site, 25 sites, and 50 sites. Solutions for problems of maintaining a one site database do not always apply at the 25 site level, and solutions at the 50 site level are unlike the levels below.

Data transmission speed is a major problem in the polling phase. Generally, the modems supplied with these systems operate at 1200 baud; while this is fast enough from a user's perspective when polling one site, it requires a large amount of "dead time" (waiting for the system to finish data collection) when polling many buildings. It was decided at the one site level to poll once per week. Currently, polling requires nearly eight hours per week to gather data from all 54 sites. Unfortunately, these data loggers are generally installed in machine rooms that contain so much electrical noise that faster data transmission rates are probably unreliable. Because sites with fewer channels can hold many more records (the amount of memory in these DAS is finite and constant over all loggers), it might seem like a good idea to poll the smaller sites less

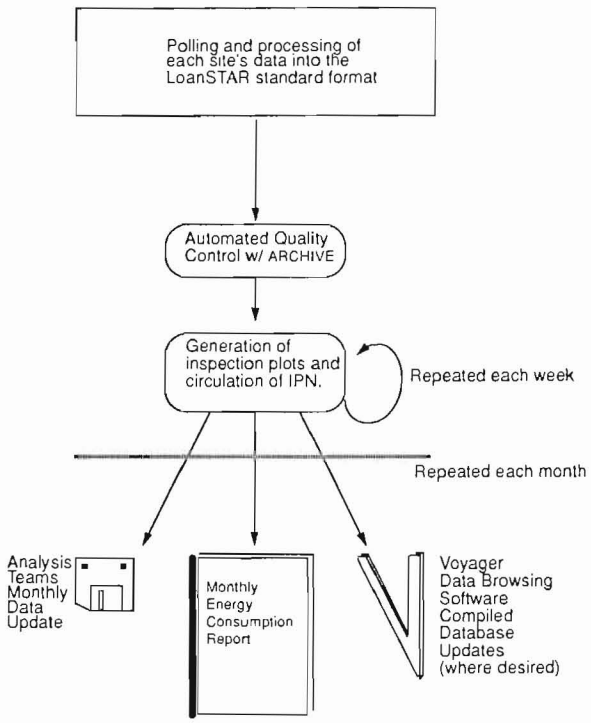

***Figure 4**: Summary of data management paths for building energy data.* This figure shows the normal pathway of data through the entire data management scheme.

frequently than once per week to alleviate the workload. However, polling every logger at a weekly interval allows malfunctioning loggers to be identified with a minimum amount of unrecoverable data. This is a primary quality control measure. Additionally, this kind of a change would require much retooling of the processing routines which depend on the weekly nature of all the data files.

Although Figure 4 gives the illusion of a simple flow between points, an extra dimension should always be envisioned: the processing routines must be performed for every site. As with the polling phase, this leads to a generous amount of idle time while waiting for the processing computer to complete the processing stream for all sites. Until August 1991, the polling and processing computer was also utilized as a general purpose machine on non-peak days. Because graduate students and staff used some of the same commercial software, such as the graphics package, for other purposes, the Database Manager was constantly on the alert to "fix leaks" in the processing stream: those created when the unknowing user has changed some important setting in the software. The willingness of the project to purchase a dedicated polling and processing computer has lessened this problem greatly. This dedicated computer is an IBM PC compatible based on an 80386-SX processor with 100 megabytes of disk space. With 54 sites, this computer requires six hours to process the collected data and create inspection plots, roughly seven minutes per site.

Although the project has been in a production phase for many months, occasionally unexpected problems occur with the data, and the routines must be retooled. For example, a particular DAS vendor recently released a new version of its polling software. The new software produces output that is just slightly different from the previous version. This caused severe problems for the first AWK filter, and consequently all subsequent routines. While simple modifications to this AWK script fixed the problem, this illustrates the fragility of a sequential processing system.

For data quality control, the circulation of the IPN is tremendously valuable. Problems with processing, metering constants, and the equipment itself are usually identified and solved by this method. Because the IPN circulates weekly, these problems can be addressed promptly, and a minimum of valuable data is lost. The production of monthly plots has also greatly improved the quality of data because the MECR allows the staff to visualize a larger window of information than the weekly plots. The logistics of circulation of the IPN and preliminary versions of the MECR between several Principal Investigators and key staff members should not be taken lightly, however. At the single site level, circulation of the inspection "notebook" was trivial. At the 25 site level, a balance had to be created in the IPN between notebook size and the time window of data which needed to appear for quality control. Four or five weeks of data has proven to be a reasonable window: much more than that seems to be too hefty a stack to check coherently. Unfortunately, four weeks of data for 25 sites is a substantial notebook to circulate. Therefore, at the 50 site level, the notebook was split into two 5-inch thick volumes. Additionally, at the 25 site level, it was reasonable for each interested person to spend four hours a week auditing the most recent data. At the 50 site level, the associated increase in review time was not linear; in fact, the time required became roughly ten hours per week. To lighten the load, the current procedure is for each person in the process to have a strict list of items to check. Also, certain members of the review committee view only one volume of the IPN each week. This does not cut the viewing time in half, because now two weeks of new information must be reviewed for the current volume, but the time is decreased.

Intertask communication among all aspects of the program has also proven essential for data quality. As new sites are mapped and evaluated for metering, immediate contact is formed with the field engineers and the installation contractor. By developing this rapport at an early stage, many problems at later stages can be reduced or eliminated entirely. In some cases, it has proven desirable to monitor new channels after a building retrofit has been completed. Adding new channels to the middle of a sequential processing scheme that utilizes columnar data can cause great problems for both the processing routines and the analysis teams. Participation in such decisions by the Database Manager at the earliest possible stage helps to reduce the stress and confusion during the transition period. Intertask communication also narrows the sleuthing required to track down and solve complex problems. As stated before, many apparent "metering problems" can actually be traced to processing glitches and fixed without expensive site visits by the field engineers. Similarly, some real metering problems can be identified at the polling stage and dealt with quickly, if the correct lines of communication have been developed.

When each site is viewed in isolation, the collection of data, quality assurance of the data, and production of reports for the building require small amounts of work, and the return on the investment is very high. The amount of time and effort required to manage data efficiently for multiple sites (more than 25) at the same rate of return has proven to be tremendous. Nuisances at the one site level become major roadblocks at the multiple site level. As a simple example, the time required to print a graph is not usually taken into account. The first laser printer that the project owned required almost five hours to print the inspection plots every week, and this was at the 25 site level. This caused severe strains on productivity for other people needing access to the printer. Maintenance costs were also increased dramatically. Even the smallest items are magnified greatly when they have to be repeated many times each week. However, the assurance of quality data as well as the rapid turnaround time of monthly reports to participating agencies are essential.

CURRENT PROJECTS

The MAP is currently investigating a reasonable format for representing LoanSTAR building data and weather data using a commercial relational database framework. While small models have been developed in the DOS environment (Haberl et al., 1991), the sheer size of the data stream has made further development much harder. Most DOS machines and database packages perform poorly given the large dataset size. The transition to an integrated database in the UNIX environment is costly: up to $10,000 for the software alone. Therefore, the project has been very cautious in this area. However, in October 1991, Informix (Informix, 1990) was installed on the

UNIX server. The benefits of using a commercial package for data management rather than simple flat files are tremendous. The man-hours required for data retrieval and delivery can be reduced substantially using a structured query language (SQL). A commercial database will allow for much higher data confidence and quality at a lower manpower cost, both in terms of development and maintenance. The representation of building data in a relational format that is efficient both in space utilization and retrieval time has proven somewhat difficult. Time series data are particularly difficult to represent with the relational model (McKenzie and Snodgrass, 1991). However, steady progress is being made.

Concurrent with the development of a relational data model in Informix, the project is currently investigating methods for producing the IPN and MECR from within Informix using SQL, C functions, and SAS. Prototype sections of the MECR have been produced already. Using the database and the multiprocess capabilities of UNIX, many of the current functions that require a dedicated PC and operator could be automatically executed on the server unattended. This would free up both machines and manpower.

Public domain software is being developed to poll particular vendors' loggers, process the data into the LoanSTAR format, and perform quality control - all in one package. This will replace many of the steps in the current data stream. In turn this should eliminate much of the damage control / problem resolution which occurs periodically. Much of the polling work will be automated by having the PC call the loggers automatically during the night to download data. This will free nearly 20 hours a week of valuable work time.

These three areas of work represent a movement away from the repetitive, task oriented nature that has embodied the data collection effort during the first two and a half years of the project. As progress is made in these areas, the focus of human and machine resource utilization can be shifted into higher level areas, such as analysis and more detailed agency feedback. However, if past experience is any indicator, new levels of complexity and new sets of problems, will probably be reached as more sites are added. The importance of a move towards a more automated system is paramount in any case.

FUTURE DIRECTIONS

More advanced methods for quality control will be investigated. Routines will be developed to scan the weekly diagnostic log files generated by ARCHIVE and other routines automatically and generate summary reports of the problems encountered at the various sites. Work is beginning on the replacement of the static upper and lower data bounds in channel tables with dynamic data limits. For example, tighter bounds can be used to check outdoor temperature data depending on the season. Also, this might allow standard relations between channels to be verified. For example, submetered electricity consumption at a building should obviously be less than the whole building consumption; at sites with numerous submetered points, human verification of this simple fact can become tedious.

The two most interesting future milestones will also require the most research. At the fifty site level, scanning the IPN each week has proven tedious at best. Auditing the data weekly for one hundred sites is probably too much to ask. Work has begun towards the development of an on-line hypertext IPN (Willis and Haberl, 1992). Soon, work will begin in earnest on investigating statistical-based pattern recognition and machine learning techniques (Michalski et al., 1983) to automatically scan the data and search for problems. This could be potentially the most exciting advance to come from the project.

CONCLUSION

The LoanSTAR MAP is producing a set of public domain software modules that can process data collected from buildings, perform data quality control checks, produce professional quality production graphs for external release, and package the data for analysis. By using off-the-shelf components and small patchwork scripts, the cost has been reduced to a bare minimum. Feedback is being given to participating agencies on a regular, timely basis. There is still much that can be done, but the project was able to go into a production mode within six months after the first new sites were added. In retrospect, many data management issues can be identified as a function of the number of sites involved; unfortunately, this function does not appear to be linear. Many of the management headaches remain, but the overall return on the investment actually seems to be very high. Several retrofits already have been significantly improved because of the feedback that the program has provided. Initial responses from participating agencies are generally favorable.

ACKNOWLEDGMENTS

The authors would like to acknowledge several people for their participation in the preparation of this paper. Dr. Srinivas Katipamula, Ms. Kristel Weber, and Mr. Dean Willis were instrumental in the early days of the project in designing prototype versions of the processing software. Mr. Satish Ramanathan is responsible for the weekly polling, processing, and basic IPN circulation. Mr. Kelly Kissock has provided generous ideas and feedback concerning the data retrieval and report generation phases. Mr. Kissock is also responsible for introducing the project to the VOYAGER data browsing program. Mr. Robert Sparks, the Programming Manager of the LoanSTAR MAP, has been responsible for much of the software involved in all phases of data management and provided valuable ideas concerning this paper. He has done the preliminary investigation of relational data models, and will lead the way into machine learning. Dr. David Claridge has provided generous assistance and guidance with the creation and format of the IPN and the MECR.

This project is funded and supported by the State of Texas, Governor's Energy Office, as part of Texas A&M University's LoanSTAR Monitoring and Analysis Program contract using oil overcharge funds.

REFERENCES

Claridge, D., Haberl, J., Sparks, R., López, R., and Kissock, K., 1992, "Monitored Commercial Building Energy Data: Reporting the Results", *ASHRAE Transactions*, in preparation.

Feuermann, D., and Kempton W., 1987, "ARCHIVE: Software for the Management of Field Data", *Center for Energy and Environmental Studies Report No. 216*, Princeton University.

FSF, 1989, GAWK, Free Software Foundation, 675 Massachusetts Ave., Cambridge, Massachusetts, 02139.

Haberl, J., Katipamula, S., Willis, D., Weber, K., Matson, J., Rayaprolu, M., and Subramanian, U., 1990a, "Task 4 Progress Report", *Texas LoanSTAR Monitoring and Analysis Program Progress Reports*, submitted to the Governor's Energy Office, State of Texas, July.

Haberl, J., Katipamula, S., Willis, D., Weber, K., Matson, J., Rayaprolu, M., and Subramanian, U., 1990b, "The Texas LoanSTAR Program: Acquiring and Archiving Data", *Proceedings of the Seventh Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, August.

Haberl, J., Jagannathan, V., López, R., Sparks, R., Kissock, K., Willis, D., and Claridge, D., 1991, "Exploring an Integrated Database Structure for Building Energy Monitoring Data", *International Building Performance Simulation Association Proceedings*, August.

Informix, 1990, INFORMIX, Informix Software Inc., 4100 Bohannon Drive, Menlo Park, California, 94025.

Knuth, D, 1986, *The TEXbook*, The American Mathematical Society and the Addison-Wesley Publishing Company, Reading, Massachusetts.

Lantern, 1990, VOYAGER, Lantern Corporation, 63 Ridgemont Drive, Clayton, Missouri, 63105 (requires Microsoft Windows).

McKenzie, L., and Snodgrass, R., 1991, "Evaluation of Relational Algebras Incorporating the Time Dimension in Databases," *ACM Computing Surveys*, Volume 23, Number 4, December.

Michalski, R., Carbonell, J., and Mitchell, T., 1983, *Machine Learning: An Artificial Intelligence Approach*, Morgan Kaufmann Publishers Inc., Los Altos, California.

SAS, 1990, Statistical Analysis Software, SAS Institute, SAS Circle, Box 8000, Cary, North Carolina.

Willis, D., and Haberl, J., 1992, "A Collaborative Support System for the Review of Building Energy Data in the LoanSTAR MAP", *Proceedings of the Eighth Annual Symposium on Improving Building Systems in Hot and Humid Climates*, Texas A&M University, May.